Правительство Российской Федерации

Государственное образовательное бюджетное учреждение высшего профессионального образования

«Национальный исследовательский университет Высшая школа экономики»

Факультет компьютерных наук
Департамент анализа данных и искусственного интеллекта

Программа дисциплины

«Анализ данных на платформе SAS»

Автор программы:
Ильвовский Д.А., (dilvovsky@hse.ru),

Москва, 2017

1. Аннотация

Дисциплина «Анализ данных на платформе SAS» предназначена для подготовки бакалавров и магистров направления «Прикладная математика и информатика». Она входит в цикл дисциплин, связанных с основами информационных технологий и анализа данных.

В курсе изучаются задачи анализа данных и методы их решения на программной платформе SAS. Основное внимание уделяется статистическим методам анализа данных. Рассматриваются основы программирования на языке SAS Base. Затрагиваются вопросы создания программ и макропрограмм. Обзорно освещаются нестатистические методы анализа данных и их реализация на платформе SAS.

Теоретический материал курса подкрепляется практическими занятиями по программированию заданий по изучаемой тематике.

Успешно прошедшие курс студенты получают сертификат от компании SAS.

2. Цели освоения дисциплины

Данная дисциплина ставит своей целью изучение базовых сведений по анализу данных в среде SAS. Эти знания и навыки необходимы в профессиональной деятельности специалистов по математическому моделированию и информатике.

3. Компетенции, формируемые в результате освоения дисциплины

В результате изучения дисциплины студенты должны:

- Знать основы языка SAS Base и уметь записывать и понимать простые программы на этом языке;
- Владеть основами макропрограммирования на языке SAS Base;
- Понимать принципы работы основных статистических методов анализа данных на платформе SAS;
- Уметь запускать и анализировать результаты выполнения основных статистических методов анализа данных на платформе SAS;
- Знать список основных методов анализа данных, реализованных на платформе SAS.

4. Тематический план дисциплины

№	Наименование разделов и дисциплин	Всего час.*	В том числе		Форма контроля
			лекции	Практические и самостоятельные занятия	
1	**Раздел 1. Аналитическая платформа SAS. Обзор технологий.**	2	*1*	*0*	
2	**Раздел 2. Язык программирования SAS/BASE**	6	*1*	*2*	Практические задания
2.1	Тема 2.1. Основы программирования на SAS/BASE		*1*	*1*	
2.2	Тема 2.2. Макросы, SQL		*0*	*1*	
3	**Раздел 3. Библиотека методов стат. Анализа SAS/STAT**	20	*4*	*6*	Практические задания
3.1	Тема 3.1. Введение в SAS/STAT, дисперсионный анализ		*1*	*1*	
3.2	Тема 3.2. Линейная регрессия		*0*	*2*	
3.3	Тема 3.3.Логистическая регрессия		*1*	*1*	
3.4	Тема 3.4. Обобщенные линейные модели		*1*	*1*	
3.5	Тема 3.5. Кластеризация		*1*	*1*	
4	**Раздел 4. Обзор современных возможностей SAS**	4	*2*	*0*	
4.1	Тема 4.1. Введение в Text Mining		*1*	*0*	
4.2	Тема 4.2. Некоторые современные модели анализа данных в SAS		*1*	*0*	
Итоговый контроль			***Экзамен***		

* 1 занятие (лекция или семинар) = 2 академических часа

5. Формы контроля знаний студентов

Дисциплина «Анализ данных на платформе SAS» читается в I семестре.

Тип контроля	Форма контроля	Параметры
Текущий контроль	Домашнее задание	Выдается для поэтапного выполнения в течение модуля
Итоговый контроль	Экзамен	Письменная работа 80 минут

Критерии оценки знаний

На текущем и итоговом контроле студент должен продемонстрировать владение основными понятиями из пройденных тем дисциплины.

Текущий контроль включает письменное задание, состоящее из нескольких задач по пройденному материалу.

Итоговый контроль проводится в форме письменного экзамена, включающего несколько вопросов и задач по темам дисциплины.

Порядок формирования оценок по дисциплине

Преподаватель оценивает самостоятельную работу студентов по выполнению домашних работ, выдаваемых на семинарских и практических занятиях. При этом оценивается правильность, эффективность и оформление программного кода. Оценки за домашние задания выставляются в рабочую ведомость, и перед экзаменом за домашние задания выставляется результирующая оценка по десятибалльной шкале $O_{сам.\ работа}$. Домашнее задание, сданное позднее объявленного срока, не оценивается.

Оценка итогового контроля выставляется по следующей формуле:

$$O_{итог} = 0{,}5 \cdot O_{экзамен} + 0{,}5 \cdot O_{сам.\ работа}.$$

и округляется до целого числа арифметическим способом,

где $O_{экзамен}$ – оценка за работу непосредственно на экзамене по десятибалльной системе.

Студенты, имеющие $O_{сам.\ работа}$ не ниже 9 баллов, автоматически получают такую же оценку за экзамен после прохождения короткого устного собеседования с преподавателем.

Сертификат вручается студентам, имеющим итоговую оценку не ниже 9 баллов.

6. Образовательные технологии

В преподавании данной дисциплины сочетаются:

- лекции в традиционной форме;
- семинарские и практические занятия, в ходе которых решаются задачи по темам курса;
- домашние практические задания по программированию по ряду основных тем дисциплины.

Методические указания по освоению дисциплины.

Для лучшего усвоения дисциплины рекомендуется:

- пройти онлайн-курсы:
  - SAS Programming I: Essentials https://support.sas.com/edu/schedules.html?ctry=us&id=277
  - SAS Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression https://support.sas.com/edu/schedules.html?ctry=us&id=1979
- читать:
  - Г.И.Ивченко, Ю.И.Медведев. Математическая статистика
  - О дисперсионном анализе в SAS/STAT:

http://support.sas.com/documentation/cdl/en/statug/66103/HTML/default/viewer.htm#statug_introanova_toc.htm

7. Учебно-методическое и информационное обеспечение дисциплины

Базовая литература – документация SAS (http://support.sas.com/documentation):

1. SAS/STAT(R) 9.3 User's Guide
2. SAS(R) 9.3 Functions and CALL Routines: Reference

Основная литература – online-курсы обучения:

3. Base SAS(R) 9.3 Procedures Guide, Second Edition
4. SAS(R) 9.3 SQL Procedure User's Guide.

Программная среда для работы:

http://www.sas.com/en_us/software/university-edition.html

Подробный перечень всех понятий и концепций, раскрываемых в каждой из тем курса, и литература по каждой теме размещены на сетевом ресурсе

(папке Dropbox), доступном всем студентам, проходящим курс, в виде лекций и дополнительных материалов.

8. Демонстрационный вариант формы текущего контроля

1. Используя процедуру MEANS, познакомиться с данными. [P1] Вывести средние в новый набор данных. [P2] Построить график процедурой SGPLOT, используя полученный набор данных: strength но оси Y, Additive по оси X, группировать по переменной Brand. [P3] Что вы можете сказать о данных? [P4] Основываясь на графике, нужно ли использовать пересечение факторов Additive и Brand в модели?

2. [P1] Проверьте гипотезу о том, что средняя прочность одинакова для всех марок. Проверить предположения. Если возможно, сравните все марки с маркой Graystone. [P2] Добавьте оставшийся фактор – Additive. Какие выводы можно сделать сейчас? [P3] Если графики из п.1 говорят, что нужно использовать пересечение, то добавьте его. Какие выводы вы можете сделать на данном шаге анализа?

3. Выполните подходящие множественные сравнения для статистически значимых переменных.

Промежуточная аттестация не предусмотрена.

9. Демонстрационный вариант экзамена

1. В исходной таблице TAB1 есть (в случайном порядке) 95% данных с входной переменной X и с целевой переменной Y, равной 1, и 5% данных с её значением, равным нулю. Создайте такую обучающую выборку TAB2 , куда попадут все наблюдения с целевой переменной, равной нулю, и столько же наблюдений с целевой переменной, равной единице (строки с единицами можно выбрать как угодно: последовательно, случайно, каждую N-ую и т. д.)

2. Пусть есть таблица work.tab1, содержащая (среди прочих) переменную a (числовую, интервальную). Напишите шаг данных для стандартизации значений в переменной a (Z-scoring), то есть их замене по формуле a=(a-mu)/sigma, где mu - выборочное мат. ожидание переменной a, a sigma - среднеквадратичное отклонение переменной a. При вычислении значений mu и sigma игнорируйте пропущенные значения.

3. Дан код:

%let a=10; data test; r= 0; run ;
Какие макроинструкции вам обязательно понадобятся, чтобы вывести "a" раз наблюдение r=0 в набор данных test :
(%for) (for)(%macro) (output) (%output) (%do) (do) (end) (%end) (%mend)

4. Выбирая значимые переменные с помощью процедуры PROC VARCLUS, ориентируются на критерий 1-R^2, потому что он позволяет:

a) выбрать предикторы, сильнее всего связанные с целевой переменной
b) оценить оптимальное количество кластеров
c) сгруппировать в один кластер несколько коррелирующих переменных
d) выбрать наиболее репрезентативный предиктор из группы коррелирующих признаков